Diagnostic accuracy of the Patient Health Questionnaire-9 for assessment of depression in type II diabetes mellitus and/or coronary heart disease in primary care

Article (Accepted Version)

Diagnostic accuracy of the Patient Health Questionnaire-9 for Assessment of Depression in Type II Diabetes Mellitus and/or Coronary Heart Disease in Primary Care-

G Lennart van der Zwaan,

Susan EM van Dijk,

Marcel C Adriaanse,

Harm WJ van Marwijk,

Maurits W van Tulder,

Alide D Pols,

Judith E Bosmans

ABSTRACT

Objective: Depression is common among type 2 diabetes mellitus (DM2)/coronary heart disease (CHD) patients and is associated with adverse health effects. A promising strategy to reduce burden of disease is to identify patients at risk for depression in order to offer indicated prevention. This study aims to assess the diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) to be used as a tool to identify high risk patients.

Methods: In this cross-sectional study, 586 consecutive DM2/CHD patients aged >18 were recruited through 23 general practices. PHQ-9 outcomes were compared to the Mini International Neuropsychiatric Interview (MINI), which was considered the reference standard. Diagnostic accuracy was evaluated for minor and major depression, comparing both sum- and algorithm based PHQ-9 scores.

Results: For minor depression, the optimal cut-off score was 8 (sensitivity 71%, specificity 71% and an AUC of 0.74). For major depression, the optimal cut-off score was 10 resulting in a sensitivity of 84%, a specificity of 82%, and an AUC of 0.88. The positive predictive value of the PHQ-9 algorithm for diagnosing minor and major depression was 25% and 33%, respectively.

Limitations: Two main limitations apply. MINI Interviewers were not blinded for PHQ-9 scores and less than 10% of all invited patients could be included in the analyses. This could have resulted in biased outcomes.

Conclusions: The PHQ-9 sum score performs well in identifying patients at high risk of minor and major depression. However, the PHQ-9 showed suboptimal results for diagnostic purposes. Therefore, it is recommended to combine the use of the PHQ-9 with further diagnostics to identify depression.

INTRODUCTION

The presence of minor depression among patients with a chronic disease is high [1, 2]. In type 2 diabetes mellitus (DM2) and/or coronary heart disease (CHD) patients for example, the 12-month prevalence of minor depression ranges between 25-40% [3, 4]. Approximately 40% of these DM2 and/or CHD patients with depressive symptoms will develop major depressive disorder within two years, indicating that the presence of depressive symptoms or minor depression is the most important predictor for major depression [3].

The yearly prevalence of major depression is comparable in DM2 patients and patients with CHD and adds up to about 10 -20 percent, which is considerably high compared to the prevalence of 5% in the general Dutch population [1, 3-10]. Major depression among DM2 and/or CHD patients is associated with lower quality of life, an increased risk of mortality, poor medication adherence and increased health care costs [9, 11-16]. Moreover, once patients are diagnosed with major depression, only roughly one third of the associated disease burden can be averted, even when optimal treatment is in place [17].

These severe consequences of depression underline the importance of correctly identifying minor and major depression in patients with DM2 or CHD thereby enabling general practitioners (GP's) to provide suitable depression care. Previous studies have shown that collaborative and stepped care programs may successfully prevent the development of major depression in patients with minor depression [18, 19]. However, minor depression often remains unrecognized, which hinders successful indicated prevention [20, 21].

Although there is no evidence supporting the routine screening of all primary care patients for depression, general practitioners should be alert to the existence of depression in patients with chronic medical conditions such as DM2 and CHD [22]. Thus, GPs have a need for instruments that can be efficiently used to identify depression in their patients and to monitor the course of depression during treatment.

Several instruments have been developed over the past years to identify and monitor depression[23] and many of them are tested in patients with chronic medical illnesses [24]. One of these instruments is the Patient Health Questionnaire-9 (PHQ-9) which is an extensively applied self-reported questionnaire comprising the nine depression items from the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) [25]. A recent meta-analysis reported good diagnostic accuracy of the PHQ-9 for the case identification of major depression in patients with chronic medical illness. This meta-analysis identified 6 studies that all used the Diagnosis and Statistical Manual of Mental Disorders (DSM) or the International Classification of Diseases (ICD) as reference standard. The pooled sensitivity and specificity were 84% and 88%, respectively, and the pooled positive and negative likelihood ratios were 6.77 and 0.19, respectively [24]. Another study, not included in this systematic review, compared the performance of the PHQ-9 to a diagnosis of major depression according to the DSM-IV criteria in clinical outpatients with DM2. Specificity was found to be good (80%-87%), but sensitivity was relatively low (58%-76%). [26]. Additionally, Hadad et al found similar results in 2013 in primary care patients with CHD (sensitivity 59-94%, specificity 84%-95%) [12].

However, studies included in the aforementioned meta-analysis were conducted in a very heterogeneous group of patients with different chronic physical health problems. Furthermore, to our knowledge, no information is available for the performance of the PHQ-9 to detect minor depression in DM2 and CHD patients in primary care, which is important because this is one of the main risk factors for major depression. Therefore, this study aims to assess the diagnostic accuracy of the PHQ-9 to identify minor and major depression in DM2 and CHD patients without known depression in primary care.

METHODS

Design

This study was conducted alongside the StepDep study; a cluster randomized controlled trial to evaluate the cost effectiveness of a stepped-care intervention to prevent major depression in DM2 and/or CHD patients in primary care. Detailed information on this project can be found elsewhere [27]. Data for purposes of the current study were collected using a cross-sectional design in a convenience sample from the StepDep study.

Ethical approval

The study-protocol was approved by the Medical Ethics Committee of the VU University Medical Centre. This study was conducted corresponding the principles of the Declaration of Helsinki and the Dutch Medical Research Involving Human Subjects Act.

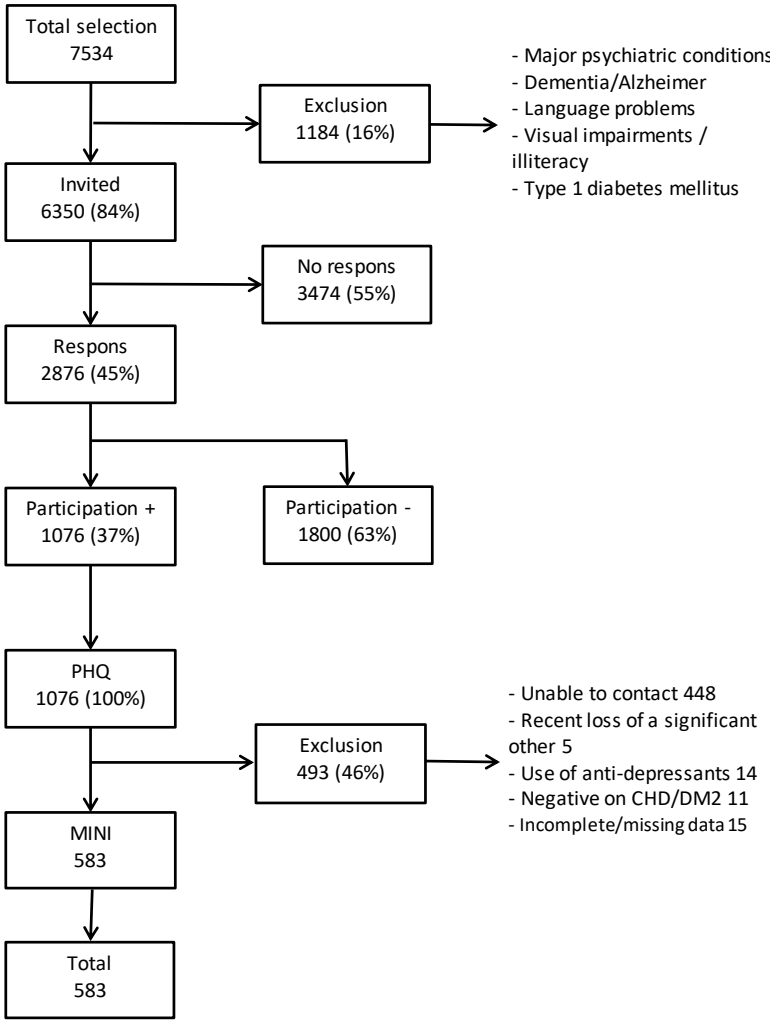
Study population

The study population was recruited through 23 general practices with 58 GPs and 128,980 enlisted patients in the vicinity of Amsterdam and Twente in the Netherlands. All GPs constructed an initial list of patients with DM2 and or CHD (angina pectoris, acute myocardial infarction, other/chronic ischemic heart disease, coronary sclerosis, previous myocardial infarction) according to the ICPC (n= 7,534) (International Classification of Primary Care, see Appendix 1) and were requested to exclude patients fulfilling the pre-set exclusion criteria. In total, 1,184 patients (16%) were excluded from participation by the GP, because they were diagnosed with type 1 diabetes mellitus, dementia/Alzheimer, major psychiatric conditions (schizophrenia, bipolar depression, affective psychosis, borderline, suicidal attempts), mental retardation, visual impairment, illiteracy, recent loss of a significant other, pregnancy, the use of anti-depressants, or not mastering the Dutch language (Appendix 1). All remaining individuals (n= 6,350) were invited by mail, through a letter from their GP, to complete the written version of the PHQ-9, which was attached to the invitation

letter. Demographical information such as gender and zip code was collected concurrently with the PHQ-9.

A total of 2,876 patients (45%) returned the PHQ-9 by mail and of those 1,076 (37%) gave informed consent to participate. These patients were approached by telephone for a MINI interview. The interviews took place within two weeks after receiving the PHQ-9 and were administered by trained interviewers. Patients who responded but did not give informed consent (n=1800) were excluded from the study. Figure 1 represents a flowchart of the recruitment of respondents. Another 448 (41%) patients were excluded because they could not be contacted within two weeks. Additionally, thirty-three were excluded of whom five recently lost a significant other, 14 used anti-depressants at time of the inclusion, eleven did not have DM2 or CHD, and 15 had incomplete/missing data. Valid scores for the final (complete-case) analysis were obtained from 583 patients.

Figure 1 - Flowchart of the study population



Measurements

PHQ-9: The PHQ-9 is a brief instrument for screening and diagnosing depressive symptoms that was developed by Kroenke et al. [8]. The continuous sum score can be used for screening purposes while the dichotomous algorithm score can be used for diagnostic purposes. The PHQ-9 comprises the following nine items which evaluate the presence of one of the nine symptoms of depression based on the DSM-IV criteria: (a) depressed mood, (b) anhedonia, (c) trouble sleeping, (d) feeling tired, (e) change in appetite or weight, (f) guilt or worthlessness, (g) trouble concentrating, (h) feeling slowed down or restless, (i) suicidal thoughts. These nine items have the following answer categories: "not at all" (0), "various days" (1), "more than half the days" (2) and "almost every day" (3), resulting in a sum score of 0 to 27. The Dutch version of the PHQ-9, which was translated and validated by Zuithof et al., was used for the current study [28].

Earlier research reported scores in the 5-9 range are considered an indication for minor depression while scores of 10 or higher were considered an indication for major depression [8, 25]. Since little is known about the most appropriate cut-off score for minor and major depression in this particular high risk population this study aims to establish cut-off scores for screening purposes for this group. Based on the PHQ-9 algorithm, minor depression is diagnosed when at least one of the essential features of major depression (depressive mood/anhedonia) and one, two or three additional depressive symptoms occurred in the previous two weeks for more than half the days [8]. Major depression is diagnosed when patients score two or more on at least five categories including one of the essential features of major depression (depressive mood/anhedonia) in the previous two weeks. The PHQ-9 outcomes in the current study were interpreted without knowledge of the results of the reference standard.

MINI: The MINI interview was used as reference standard to diagnose depression. The MINI is a short structured diagnostic interview to diagnose DSM IV disorders that can be administered by non-specialized interviewers. Furthermore, the MINI is often used in clinical practice because of its short

administration time of approximately 15 minutes. Various studies reported positive results regarding the validation of the MINI, with high sensitivity (0.70 or higher) and specificity (0.85 or higher) [29, 30].

Analysis

Both instruments, the PHQ-9 and the MINI, are based on DSM IV criteria and the outcomes will be compared in order to assess diagnostic accuracy. Diagnostic accuracy was established in terms of: sensitivity/specificity; positive/negative predictive values, likelihood ratios and Receiver Operating Characteristic (ROC) curve analysis.

PHQ-9 sum score: Operating characteristics for the continuous sum scores, applied for screening purposes, were calculated by conducting ROC curve analysis. The ROC-curve summarizes the true positive fraction (sensitivity) on the y-axis as a function of the false positive fraction (1-specificity) on the x-axis for diverse cut-off scores. In addition, the area under the curve (AUC) summarizes the tests ability in actually discriminating individuals with and without minor/major depression. A maximum score of 1 indicates a perfect distinction between the diseased (true-positive fraction = 100) and the non-diseased (false negative fraction = 0). A rough guide for interpreting the AUC and classifying the accuracy of a diagnostic test is as follows: 0.9 – 1 = excellent; 0.8 – 0.9 = good; 0.7 – 0.8 = fair; 0.6 – 0.7 = poor; 0.5 – 0.6 = fail [31, 32]. In this study, the AUC was calculated based on the sum-scores of the PHQ-9.

PHQ-9 algorithm: Test characteristics for diagnostic purposes were calculated using the dichotomous algorithm score (Appendix 2). The sensitivity of a test indicates the percentage of people who were correctly classified by the test as having the disorder. Specificity refers to the percentage of people who were correctly classified by the test as not having the disorder. The likelihood ratio is a measure that reveals the likelihood of a positive or negative outcome. The positive likelihood ratio (LR +)

indicates the extent to which a disease is more likely in a patient after finding a positive test result. The negative likelihood ratio (LR-) indicates the extent to which a disease is less likely with a negative test result. Although these test-characteristics may be valuable in clinical decision-making, the prior probability of the disease must also be taken into account. A high prevalence will increase the positive predictive value (PPV) by influencing the true positive and false positive rates. The negative predictive value (NPV), on the contrary, decreases with high prevalences [33]. Therefore, predictive values were also estimated. The positive predictive value is the percentage of people with a positive test outcome who actually have the disease. The negative predictive value on the contrary is the percentage of people with a negative test outcome who do not have the disease [34].

SPSS version 21.0 was used to calculate various test characteristics, to plot the corresponding ROC-curve and to assess the optimal cut-off point by calculating the AUC. As patients who are diagnosed with major depression also fulfil the criteria for a minor depression, results are presented for the combined diagnosis of minor or major depression and for major depression alone.

RESULTS

Sample characteristics

The mean age of the participating patients was 68.9 years (SD = 9.8), with 61.8% of the sample being male. The mean PHQ-9 score was 5.76 (SD = 4.8). Participating practices were located in four different locations, namely; Amsterdam (metropolitan area), Hengelo (urban area), Delden and Borne (rural area). Baseline characteristics are presented in Table 1.

Table 1 - Baseline characteristics

Characteristic	N = 583
Age (Mean, sd)	
Years	68.9 (9.8)
Gender (n=, %)	
Female	223 (38.2%)
Male	360 (61.8%)
PHQ-9 score (mean, sd)	
Range 0-27	5.76 (4.8)
General practice (n=, %)	
Amsterdam	315 (54.3%)
Borne	42 (7.2%)
Delden	23 (3.9%)
Hengelo	203 (34.6%)

Operating characteristics for sum-scores of the PHQ-9

Minor or major depression: Increasing PHQ-9 cut-off scores were associated with decreasing sensitivity in diagnosing minor or major depression, while specificity increased (Figure 2). Similar results were found for the positive likelihood ratios, as higher PHQ-9 cut-off scores improved the likelihood of a minor or major depression. The negative likelihood decreased with higher cut-off scores. The most optimal cut-off score for screening purposes was 8 and resulted in a sensitivity of 0.71 (95% CI: 0.52-0.84), a specificity of 0.71 (95% CI: 0.67-0.75), a positive predictive value of 0.13 (95% CI: 0.09-0.19), a negative predictive value of 0.98 (95% CI: 0.95-0.99), a positive likelihood ratio of 2.44 (95% CI: 1.89-3.14) and a negative likelihood ratio of 0.41 (95% CI: 0.25-0.70) (Table 2).

Table 2 - Test characteristics of the PHQ-9 cut-off scores for diagnosing Minor Depression

Cut-off score	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)
≥5	0.82 (0.65-0.93)	0.49 (0.44-0.53)	0.09 (0.06-0.13)	0.98 (0.95-0.99)	1.60 (1.34-1.90)	0.36 (0.18-0.76)
≥6	0.77 (0.58-0.89)	0.55 (0.51-0.59)	0.10 (0.06-0.14)	0.97 (0.95-0.99)	1.70 (1.38-2.09)	0.43 (0.23-0.79)
≥7	0.71 (0.52-0.84)	0.65 (0.61-0.69)	0.11 (0.07-0.16)	0.97 (0.95-0.99)	2.02 (1.58-2.58)	0.45 (0.27-0.76)
≥8	0.71 (0.52-0.84)	0.71 (0.67-0.75)	0.13 (0.09-0.19)	0.98 (0.95-0.99)	2.44 (1.89-3.14)	0.41 (0.25-0.70)
≥9	0.59 (0.41-0.75)	0.76 (0.72-0.79)	0.13 (0.08-0.20)	0.97 (0.94-0.98)	2.45 (1.78-3.36)	0.62 (0.36-0.81)

(Abbreviations: CI, confidence interval lower limit-upper limit; PPV, positive predictive value; NPV, negative predictive value; LR, likelihood ratio)

The area under the curve for the PHQ-9 sum-score when screening for minor or major depression was 0.74 (95% CI: 0.67-0.81) suggesting a fair/acceptable test in discriminating between persons with and without a diagnosis of minor or major depression (Figure 2) [31].

Major depression: Results regarding major depression were similar to the results for minor or major depression: the sensitivity decreased for higher PHQ-9 scores as the specificity improved and the likelihood of a major depression increased similar to higher PHQ-9 scores (Figure 3). The optimal cut-off score for screening a major depressive disorder was 10 with a sensitivity of 0.84 (95% CI: 0.63-0.95) and a specificity of 0.82 (95% CI: 0.78-0.85) (Table 3).

Table 3 - Test characteristics of the PHQ-9 cut-off scores for diagnosing Major Depression

Cut-off score	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Positive LR (95% CI)	Negative LR (95%CI)
≥8	0.92 (0.72-0.99)	0.71 (0.67-0.75)	0.13 (0.08-0.18)	1 (0.98-1)	3.21 (2.69-3.82)	0.11 (0.03-0.42)
≥9	0.88 (0.68-0.97)	0.77 (0.73-0.80)	0.14 (0.09-0.21)	0.99 (0.98-1)	3.78 (3.07-4.65)	0.16 (0.05-0.45)
≥10	0.84 (0.63-0.95)	0.82 (0.78-0.85)	0.17 (0.11-0.25)	0.99 (0.98-1)	4.55 (3.56-5.81)	0.20 (0.08-0.48)
≥11	0.80 (0.59-0.92)	0.85 (0.81-0.87)	0.19 (0.12-0.28)	0.99 (0.97-1)	5.19 (3.94-6.84)	0.24 (0.11-0.52)
≥12	0.76 (0.54-0.90)	0.89 (0.86-0.91)	0.23 (0.15-0.34)	0.99 (0.97-1)	6.84 (4.96-9.44)	0.27 (0.13-0.54)

(Abbreviations: CI, confidence interval lower limit-upper limit; PPV, positive predictive value; NPV, negative predictive value; LR, likelihood ratio)

The area under the curve was 0.88 (95% CI: 0.78-0.98) indicating the sum score of the PHQ-9 is a good test for discriminating persons with and without major depression (Figure 3).

Figures 2 and 3 - ROC curves for PHQ-9 summed-score versus MINI outcome for minor or major depression and major depression *(diagonal segments are produced by ties)*

Figure 2, Minor or major depression

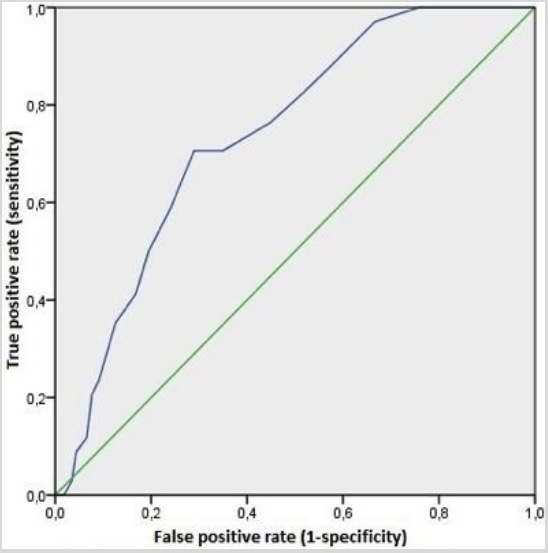
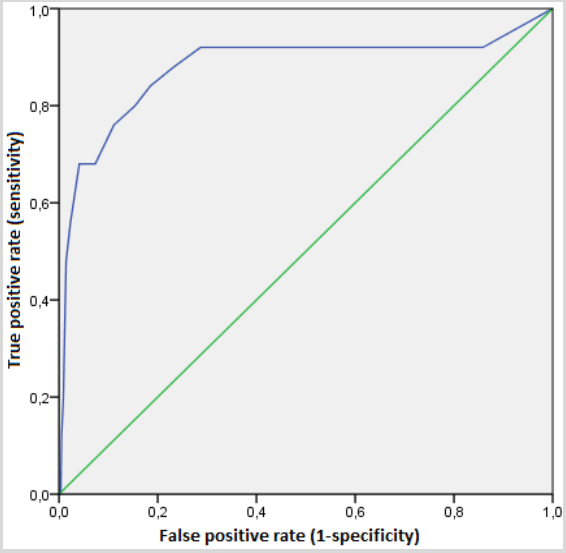


Figure 3, Major depression



Test characteristics based on the PHQ-9 algorithm

Minor or major depression: In total, 34 (5.8%) individuals were identified as having minor or major depression according to the MINI, of whom 13 (38%) were diagnosed as being depressed using the PHQ-9 algorithm (Table 4), leading to a sensitivity of 0.38 (95% CI: 0.23-0.56). Of the 549 persons without minor or major depression according to the MINI, 510 (87%) had a negative PHQ-9 algorithm outcome, which resulted in a specificity of 0.93 (95% CI: 0.90-0.95), a positive predictive value of 0.25 (95% CI: (0.15-0.39), a negative predictive value of 0.96 (95% CI: 0.94-0.97), a positive likelihood ratio of 5.38 (95% CI: 3.19-.9.08) and a negative likelihood ratio of 0.67 (95% CI: 0.51-0.87).

Major depression: A total of 25 (4.3%) patients had major depression according to the MINI, while 43(7.4%) patients were identified as having major depression based on the PHQ-9 algorithm (Table 4, bold). This resulted in a sensitivity of 0.56 (95% CI: 0.35-0.75), a specificity of 0.95 (95% CI: 0.93-0.96), a positive predictive value of 0.33 (95% CI: 0.20-0.49), a negative predictive value of 0.98 (95% CI: 0.96-0.99), a positive likelihood ratio of10.78 (95% CI: 6.56-17.70) and a negative likelihood ratio of 0.46 (95% CI: 0.30-0.72).

Table 4 - crosstable outcomes PHQ-9 algorithm versus MINI

	MINI positive		**MINI negative**		**Total**	
	Minor or Major depression	**Major depression**	**Minor or Major depression**	**Major depression**	**Minor or Major depression**	**Major depression**
PHQ-9 positive	13	14	39	29	52	43
PHQ-9 negative	21	11	510	529	531	540
Total	34	25	549	558	583	583

In Appendix 3A and 3B, all 2 by 2 classification tables are provided for each algorithm and cut-off score on the PHQ-9.

DISCUSSION

This study showed that the PHQ-9 sum-score is a valuable tool to identify patients at high risk for minor or major depression. The optimal cut-off score for minor or major depression was 8, while the most appropriate cut-off score for major depression was 10. However, applying the PHQ-9 algorithm for diagnostic purposes resulted in high false positive rates. Thus, a large percentage of respondents will be erroneously labeled as having minor or major depression. This is in line with results from previous studies evaluating the criterion validity of the PHQ-9 in chronically ill and high risk populations [13, 35-37].

Previous studies recommended to use a cut-off score of 6 based on a combination of sensitivity and specificity when screening for minor or major depression [13, 27] which is substantially lower than the optimal cutoff of 8 in our study. A possible explanation for the higher cut-off score we found may be due to overlapping symptomatology between DM2/CHD and depression. Thus, DM2/CHD related complaints, such as fatigue, are labelled as depressive symptoms and a higher cut-off score is then needed to discriminate between individuals with and without depression. Since lower cut-off scores are normally used when screening for minor or major depression in patients with DM2 and/or CHD, it is quite possible that the prevalence of minor depression is overestimated in this patient group [13, 27]. The optimal cutoff of 10 for major depression found in this study is in line with previous studies evaluating the diagnostic accuracy of the PHQ-9 for major depression in different populations [12, 26, 37]. However, this optimal cut-off score for major depression approaches the cut off score for minor or major depression (PHQ-score of 8 or more) in this patient group. This raises the question whether the PHQ-9 is specific enough to distinguish minor from major depression in this patient group.

When screening for depression it is important that the number of false positives is relatively low, meaning that specificity is more important than sensitivity. However, it is also important to correctly identify patients who are in need of care detect a substantial number of new cases, so the instrument should also be adequately sensitive [22]. As compared to other depression instruments

included in a meta-analysis of diagnostic accuracy studies in patients with chronic physical illness, the performance of the PHQ-9 in this study was better than for example the Geriatric Depression Scale (GDS), General Health Questionnaire (GHQ) and the Hospital Anxiety and Depression Scale (HADS), but similar to the PHQ-9 and Beck Depression Inventory (BDI) [24].

This study was the first that examined the performance of the PHQ-9 in screening for and diagnosing both major- and minor depression among DM2 and/or CHD patients in primary care. Second, The MINI interview, which was used as reference standard, is an extensively applied and validated instrument for diagnosing depression and can, therefore, be considered a good reference standard. Another strength is that we managed to include a large number of patients in comparison with previous studies [26, 35]. Finally, this study examined both the sum score and algorithm based PHQ-9 score which allowed us to make a comprehensive statement about the applicability of the PHQ-9 in general practice.

There are several limitations that should be considered when interpreting our results. Firstly, less than 10% of the patients who were invited to participate could be included in the analyses. This might have resulted in biased outcomes. Possibly, we would have been able to include more patients when a longer period between the PHQ-9 and the MINI administration was allowed. However, this was considered undesirable because depressive symptoms may have improved or worsened during this period which may influence the estimates of sensitivity and specificity as has been shown in an earlier study using the PRIME-MD which is comparable to the MINI [38]. Also, no detailed demographic and/or clinical information was available on non-responders, due to the ethical and privacy restraints. As a consequence we were unable to check for possible selection bias so the reported findings cannot be extrapolated to the general population. Although a large percentage of patients was not included and limited information was available for the non-responders, the aim of this study was to assess the performance of the PHQ-9 in identifying depressive symptoms in a high risk population consisting of DM2/CHD patients and for this purpose we consider our findings relevant.

Secondly the MINI-interviewers were not blinded for the PHQ-9 scores and this knowledge may have resulted in biased outcomes and, thereby, in an overestimation of sensitivity and specificity. Thirdly, the MINI interview assesses the presence of DSM-IV criteria for depression and is based on self-reported symptoms. Although the MINI is considered a good reference standard, it cannot be considered a gold standard to diagnose depression. Fourth, the StepDep study on which the current sample was based, was not powered to assess diagnostic accuracy. In this convenience sample, only 25 cases of major depression as diagnosed by the MINI were identified, which has led to imprecise results as indicated by the wide confidence intervals surrounding the estimates. Finally, because of privacy legislation in the Netherlands it was unknown whether patients had received a diagnosis or received non-pharmaceutical treatment for depression in the past. This could have biased the results. Furthermore, information concerning diagnosis of DM2 and/or CHD was not accessible for all patients in this study. Therefore it was not possible to present the results separately for DM2, CHD and DM2/CHD. However, patients diagnosed with DM2 and/or CHD are regularly seen by their GP in the context of cardiovascular risk management and, thus, recommendations regarding the use of the PHQ-9 to identify depression in this high-risk population are of practical use to them.

Conclusions

Our results show that the continuous sum-score of the PHQ-9 in general practice can be properly used to identify DM2/CHD patients at high risk of minor- or major depression, but results in a considerable number of patients with a false-positive result. Therefore, we recommend that in usual practice the PHQ-9 is followed by a formal diagnostic procedure like a diagnostic psychiatric interview to establish the presence of minor or major depression. When resources are lacking to perform further diagnostics, it is advised to only offer indicated prevention to patients who explicitly express disease burden from depressive symptoms and a need for treatment.

Role of funding source

Funding by ZonMw, the Netherlands Organization for Health Research and Development (project number 80-82310-97-12110) was obtained. ZonMw did not have any role in the study design, the collection-, analyses- and interpretation of the data, writing of the report or the decision to submit this report for publication.

Conflicting interests

The authors declare that they have no conflicting interests.

Acknowledgements

We would like to thank our interviewers, Anna Janssen, Marleen Reuser and Wendy Kersten for their contribution in collecting the data for this study. In addition, we would like to thank ZonMw, our funding organization, participating general practices and all participating patients for enabling us to perform this study.

REFERENCES

1. Barth, J., Schumacher, M., Herrmann-Lingen, C., 2004. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. *Psychosomatic medicine* 66, 802-813.

2. Harter, M., Baumeister, H., Reuter, K., Jacobi, F., Hofler, M., Bengel, J., Wittchen, H.U., 2007. Increased 12-month prevalence rates of mental disorders in patients with chronic somatic diseases. Psychotherapy and psychosomatics 76, 354-360.

3. Bot, M., Pouwer, F., Ormel, J., Slaets, J.P., de Jonge, P., 2010. Predictors of incident major depression in diabetic outpatients with subthreshold depression. Diabetic medicine : a journal of the British Diabetic Association 27, 1295-1301.

4. Thombs, B.D., Bass, E.B., Ford, D.E., Stewart, K.J., Tsilidis, K.K., Patel, U., Fauerbach, J.A., Bush, D.E., Ziegelstein, R.C., 2006. Prevalence of depression in survivors of acute myocardial infarction. Journal of general internal medicine 21, 30-38.

5. Anderson, R.J., Freedland, K.E., Clouse, R.E., Lustman, P.J., 2001. The prevalence of comorbid depression in adults with diabetes: a meta-analysis. Diabetes care 24, 1069-1078.

6. Backenstrass, M., Frank, A., Joest, K., Hingmann, S., Mundt, C., Kronmuller, K.T., 2006. A comparative study of nonspecific depressive symptoms and minor depression regarding functional impairment and associated characteristics in primary care. Comprehensive psychiatry 47, 35-41.

7. de Graaf, R., ten Have, M., van Gool, C., van Dorsselaer, S., 2012. Prevalence of mental disorders and trends from 1996 to 2009. Results from the Netherlands Mental Health Survey and Incidence Study-2. Social psychiatry and psychiatric epidemiology 47, 203-213.

8. Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine 16, 606-613.

9. Rodriguez, M.R., Nuevo, R., Chatterji, S., Ayuso-Mateos, J.L., 2012. Definitions and factors associated with subthreshold depressive conditions: a systematic review. BMC psychiatry 12, 181.

10. Rudisch, B., Nemeroff, C.B., 2003. Epidemiology of comorbid coronary artery disease and depression. Biological psychiatry 54, 227-240.

11. Ciechanowski, P.S., Katon, W.J., Russo, J.E., 2000. Depression and diabetes: impact of depressive symptoms on adherence, function, and costs. Archives of internal medicine 160, 3278-3285.

12. Haddad, M., Walters, P., Phillips, R., Tsakok, J., Williams, P., Mann, A., Tylee, A., 2013. Detecting depression in patients with coronary heart disease: a diagnostic evaluation of the PHQ-9 and HADS-D in primary care, findings from the UPBEAT-UK study. PloS one 8, e78493.

13. Lamers, F., Jonkers, C.C., Bosma, H., Penninx, B.W., Knottnerus, J.A., van Eijk, J.T., 2008. Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. Journal of clinical epidemiology 61, 679-687.

14. Lin, E.H., Heckbert, S.R., Rutter, C.M., Katon, W.J., Ciechanowski, P., Ludman, E.J., Oliver, M., Young, B.A., McCulloch, D.K., Von Korff, M., 2009. Depression and increased mortality in diabetes: unexpected causes of death. Annals of family medicine 7, 414-421.

15. Simon, G.E., 1992. Psychiatric disorder and functional somatic symptoms as predictors of health care use. Psychiatric medicine 10, 49-59.

16. Stafford, L., Berk, M., Jackson, H.J., 2007. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. General hospital psychiatry 29, 417-424.

17. Andrews, G., Issakidis, C., Sanderson, K., Corry, J., Lapsley, H. Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment of ten mental disorders**.** *Br. J. Psychiatry*. 2004;184:526–533.

18. Cuijpers, P., Van Straten, A., Smit, F., Mihalopoulos, C., Beekman , A., 2008. Preventing the onset of depressive disorders: a meta-analytic review of psychological interventions. American journal of psychiatry 165, 1272-1280.

19. Muñoz, R.F., Cuijpers, P., Smit, F., Barrera, A.Z., Leykin, Y., 2010. Prevention of major depression. Annual review of clinical psychology 6, 181-212.

20. Culpepper, L., 2012. Does screening for depression in primary care improve outcome? Current psychiatry reports 14, 345-352.

21. Gilbody, S., Sheldon, T., House, A., 2008. Screening and case-finding instruments for depression: a meta-analysis. CMAJ : Canadian Medical Association journal 178, 997-1003.

22. Thombs, B.D., Ziegelstein, R.C. Does depression screening improve depression outcomes in primary care?. *BMJ*. 2014;348.

23. Lowe, B., Kroenke, K., Herzog, W., Grafe, K., 2004. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). Journal of affective disorders 81, 61-66.

24. Meader, N., Mitchell, A.J., Chew-Graham, C., Goldberg, D., Rizzo, M., Bird, V., Kessler, D., Packham, J., Haddad, M., Pilling, S., (2011). Case identification ofdepression in patients with chronic physical health problems. British journal of general practice 61, e808-e820.

25. Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 32, 509-515.

26. van Steenbergen-Weijenburg, K.M., de Vroege, L., Ploeger, R.R., Brals, J.W., Vloedbeld, M.G., Veneman, T.F., Hakkaart-van Roijen, L., Rutten, F.F., Beekman, A.T., van der Feltz-Cornelis, C.M., 2010. Validation of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized outpatient clinics. BMC health services research 10, 235.

27. van Dijk, S.E., Pols, A.D., Adriaanse, M.C., Bosmans, J.E., Elders, P.J., van Marwijk, H.W., van Tulder, M.W., 2013. Cost-effectiveness of a stepped-care intervention to prevent major depression in patients with type 2 diabetes mellitus and/or coronary heart disease and subthreshold depression: design of a cluster-randomized controlled trial. BMC psychiatry 13, 128.

28. Zuithoff et al., 2010Zuithoff, N., Vergouwe, Y., King, M., Nazareth, I., van Wezep, M., Moons, K., Geerlings, M. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross sectional study. *BMC Family Practice*. 2010;11:98.

29. Amorim, P., Lecrubier, Y., Weiller, E., Hergueta, T., Sheehan, D., 1998. DSM-IH-R Psychotic Disorders: procedural validity of the Mini International Neuropsychiatric Interview (MINI). Concordance and causes for discordance with the CIDI. European psychiatry : the journal of the Association of European Psychiatrists 13, 26-34.

30. Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. The Journal of clinical psychiatry 59 Suppl 20, 22-33;quiz 34-57.

31. Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29-36.

32. Roelen, C.A., Bultmann, U., van Rhenen, W., van der Klink, J.J., Twisk, J.W., Heymans, M.W., 2013. External validation of two prediction models identifying employees at risk of high sickness absence: cohort study with 1-year follow-up. BMC public health 13, 105.

33. Hajian-Tilaki, K., 2013. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian journal of internal medicine 4, 627-635.

34. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., Thomas, R., 2008. Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology 56, 45-50.

35. Chagas, M.H., Tumas, V., Rodrigues, G.R., Machado-de-Sousa, J.P., Filho, A.S., Hallak, J.E., Crippa, J.A., 2013. Validation and internal consistency of Patient Health Questionnaire-9 for major depression in Parkinson's disease. Age and ageing 42, 645-649.

36. Randall, J.M., Voth, R., Burnett, E., Bazhenova, L., Bardwell, W.A., 2013. Clinic-based depression screening in lung cancer patients using the PHQ-2 and PHQ-9 depression questionnaires: a pilot study. Supportive

care in cancer : official journal of the Multinational Association of Supportive Care in Cancer 21, 1503-1507.

37. Wittkampf, K., van Ravesteijn, H., Baas, K., van de Hoogen, H., Schene, A., Bindels, P., Lucassen, P., van de Lisdonk, E., van Weert, H. The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen. Hosp. Psychiatry*. 2009;31:451–459.

38. Bakker, I.M., Terluin, B., van Marwijk, H.W., van Mechelen, W., Stalman, W.A., 2009. Test-retest reliability of the PRIME-MD: limitations in diagnosing mental disorders in primary care. European journal of public health 19, 303-307.